基于 CNKI 的论文下载频次和被引频次的相关性分析
——以《西南交通大学学报》为例

摘　要：目的：为《西南交通大学学报》学术影响力的提高提供理论依据。方法：利用中国知网引文数据库，以西南交通大学学报 2014—2018 年发表的 818 篇文献为统计数据源，分别对文献下载频次与被引频次进行数据相关性分析，计算出两者的线性相关系数，并进行了分析和探讨，讨论利用下载频次预测被引频次的可行性。结果：被引频次 TOP50 论文主要集中在 2014—2015 年（论文发表后的 4~5 年），其中，以 2014 年居首（27 篇，占比 54%）；2015 年发表的论文数其次（15 篇，占比 30%）；2017 年和 2018 年（论文发表后的 1~2 年）发表的论文都没有进入 TOP50。下载频次 TOP50 的论文集中在 2014—2016 年（论文发表后的 3~5 年），分别占比 32%。结论：下载频次在 500~1000 次时和被引频次拟合度好，高于 1000 次时分布比较分散；被引频次 20 次左右时和下载频次拟合度好，大于 20 次时分布比较分散，拟合度差。在特定范围内利用下载频次预测被引频次是可行的。

关键词：下载频次；被引频次；拟合度；相关性
中图分类号：G237.9　　文献标识码：A
文章编号：1671-0134（2019）12-109-04　　DOI：10.19483/j.cnki.11-4653/n.2019.12.033
本文著录格式：徐萍 . 基于 CNKI 的论文下载频次和被引频次的相关性分析——以《西南交通大学学报》为例 [J]. 中国传媒科技，2019（12）：109-112.

文 / 徐萍

引言

随着互联网技术的高速发展和普及，数字化成为科学论文传播的一条重要途径，文献被引频次、下载频次、期刊影响因子、CI 指数等指标在评价学术成果影响力等方面发挥了重要的作用。下载频次与被引频次作为文献价值的集中体现方式一直受到学界的广泛关注，有研究提出了利用下载频次代替被引频次来解决引文评价中时间的滞后性问题。[1-4] 因此，下载频次有望成为一个新的考量论文传播和影响的指标。

以往的研究发现 [5-12]：下载频次与被引频次（两年后的）具有较强的相关性，文献的被引频次可以通过对应的下载频次进行预测；可以用期刊的前期下载频次借助回归方程对其后期（两年）的被引频次进行估计；期刊的下载频次与被引频次高度正相关，且不是单纯的线性相关关系。也有不少研究直接对被引频次与下载频次具有高度一致性这一说法进行否认，甚至否认两者具有相关性：Andrew 等对 International Journal of Cardiology 上 2007 年下载频次和被引频次最高的前 25 篇论文进行比较分析发现，被引频次与下载频次并无明显的关系，由此对“将被引频次作为评价论文影响力的决定性因素”提出了质疑；[13] 杜秀杰等采用简单线性相关系数判断学术论文的被引频次并不是简单的与下载频次成正比。[14]

如果两者具有较强的相关性，是否直接可以利用下载频次对被引频次进行预测呢？有必要对两者关系展开进一步深入探讨。

1. 研究对象和方法

1.1 研究对象

资料来源：利用中国知网“中国引文数据库”（http://ref.cnki.net.knsref/index.aspx）的高级检索功能（检索日期为 2019 年 3 月 1 日），在“检索项”选择“刊名”，“检索词”输入“西南交通大学学报”，发布时间从 2014—2018 年的 EI 来源期刊，检索学报 2014—2018 年刊登论文的下载和被引用情况，分别按下载频次和被引频次从大到小进行排序，将文章名、发表年份、下载频次、被引频次等相关信息导入到 excel 表格中进行数据分析。

1.2 研究方法

文章的下载频次和被引频次是计量学评价体系中的两个基本指标，在研究两个随机变量间的相互关系时，常使用数理统计学中的简单线性相关系数，具体公式为 [15, 16]

$$r=\frac{n\sum_{i=1}^{n}x_i y_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{\sqrt{n\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2}\sqrt{n\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2}}$$

式中：r 为简单线性相关系数；n 为取样数量；x_i 为被引用频次；y_i 为下载频次。

基于定量分析，探讨下载频次与被引频次之间的相关性，利用相关系数判别数据的相关性大小，在分析相关性的基础上，使用曲线估计初步确定下载频次与被引频次的函数关系曲线方程，主要是选出最佳的曲线模型，对下载频次与被引频次进行拟合。

2. 研究结果

2. 1 下载频次前 50 和被引频次前 50 论文的年份分布情况

表 1　被引频次和下载频次 TOP50 论文数量分布情况　　篇

	2014 年	2015 年	2016 年	2017 年	2018 年
被引频次	27	15	8	0	0
下载频次	17	15	17	1	0

对《西南交通大学学报》2014—2018 年在知网上被引频次、下载频次 TOP50 论文的发表年份进行统计，结果如表 1 所示。由表 1 可知：被引频次 TOP50 论文主要集中在 2014—2016 年，其中，以 2014 年最多（27 篇，占比 54%）；2015 年发表的论文数其次（15 篇，占比 30%）；2017 年和 2018 年发表的论文都没有进入 TOP50。下载频次 TOP50 的论文集中在 2014—2016 年，分别占比 32% 左右，下载频次 TOP50 的论文主要集中在论文发表后的 3~5 年。

2. 2 下载频次 TOP25 论文对应具体被引频次分布情况

下载频次可反映上网论文的扩散速率，是网络环境下期刊传播效率的一个新指标，论文下载量在第一时间反映了文献的使用情况，在一定程度上反映了被使用但最终未被引用文献的价值，相对论文被引频次、下载频次具有更强的时效性。

对《西南交通大学学报》2014—2018 年在知网上下载频次 TOP25 论文进行统计分析，结果如表 2 所示。下载频次最多的为 2015 年发表的论文（4347 次），对应的被引频次为 121 次，其中，期刊论文和硕士论文的引用居多。下载频次是 2016 年发表的排第二论文的 2 倍多。

表 2　下载前 25 论文被引情况分析

序号	发表年份	下载频次	被引频次	期刊论文	博士论文	硕士论文	会议论文
1	2015 年	4347	121	71	3	44	3
2	2016 年	2037	55	21	4	30	0
3	2014 年	1807	23	7	2	14	0
4	2014 年	1459	152	65	4	74	8
5	2015 年	1237	55	27	2	25	1
6	2016 年	1131	19	13	0	6	0
7	2014 年	935	44	21	0	22	1
8	2015 年	838	8	3	1	4	0
9	2016 年	825	17	8	1	8	0
10	2015 年	824	24	13	2	8	1
11	2016 年	803	12	5	2	5	0
12	2015 年	714	21	9	2	9	1
13	2014 年	690	27	14	3	10	0
14	2016 年	680	30	13	4	12	1
15	2016 年	679	20	12	1	7	0
16	2014 年	678	55	23	4	25	3
17	2014 年	674	19	11	2	6	0
18	2016 年	665	26	9	3	14	0
19	2018 年	627	1	1	0	0	0
20	2016 年	569	14	6	1	6	1
21	2015 年	542	17	11	1	5	0
22	2015 年	534	26	14	2	8	2
23	2015 年	527	13	10	1	2	0
24	2015 年	503	13	6	0	7	0
25	2014 年	501	42	22	19	0	1

下载频次 TOP25 论文被引频次和下载频次相关性见图 1。相关系数 R=0.431，关系式为 $y_i=0.028x_i+6.3127$。由图 1 可知：下载频次主要集中在 500~1000 的范围内，对应的被引频次主要集中在 60 次以下，下载频次在 500~1000 次时和被引频次拟合度好，高于 1000 次时分布比较分散。

具体被引分布情况见图 2。由图 2 可知：下载频次 TOP25 的论文总被引频次为 854，其中，期刊论文引用 412 次，占比 48.24%；博士论文引用 64 次，占比 7.49%；硕士论文引用 351 次，占比 41.10%；会议论文引用 23 次，占比 2.69%。

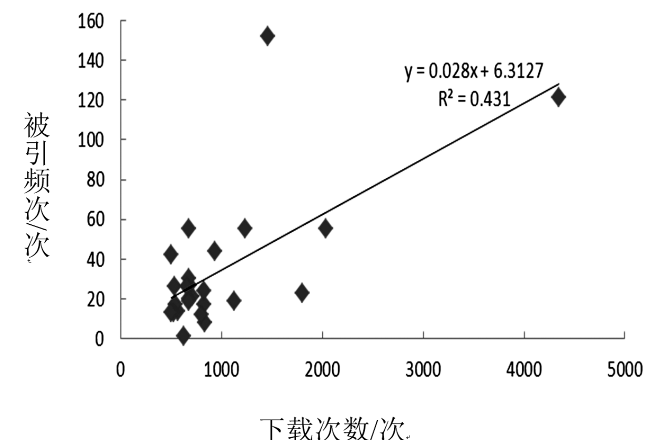


图 1　下载频次 TOP25 论文被引频次和下载频次相关性

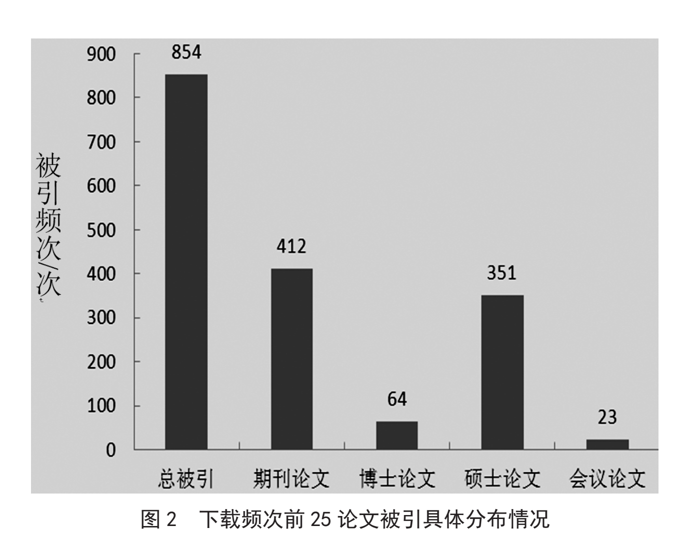


图 2　下载频次前 25 论文被引具体分布情况

2.3 被引频次前 25 论文对应下载频次及具体被引频次分布情况

对《西南交通大学学报》2014—2018 年在知网上被引频次 TOP25 论文进行统计分析，结果如表 3 所示。被引频次最多的为 2014 年发表的论文（152 次），对应的下载频次为 1459 次，其中，期刊论文和硕士论文的引用居多。被引频次与 2015 年发表的排第二的论文相差 31 次，但下载频次却相差了 3000 次左右。被引频次 TOP25 的论文的被引频次次数主要集中在 20 次左右。

被引频次 TOP25 论文被引频次和下载频次相关性见图 3。相关系数 R=0.458 3，关系式为 y_i=18.166x_i+193.1。具体被引分布情况见图 4。由图 4 可知：被引频次主要集中在 20 次左右，对应的下载频次主要集中在 500 次左右，被引频次 20 次左右时和下载频次拟合度好，大于 20 次时分布比较分散，拟合度差。

表 3　被引频次前 25 具体数据分析

序号	发表日期	被引频次	下载频次	期刊论文	博士论文	硕士论文	会议论文
1	2014 年	152	1459	65	4	75	8
2	2015 年	121	4347	71	3	44	3
3	2016 年	55	2037	21	4	30	0
4	2015 年	55	1237	37	2	25	1
5	2014 年	55	678	23	4	25	3
6	2014 年	44	935	21	0	22	1
7	2014 年	42	501	22	0	19	1
8	2016 年	30	680	13	4	12	1
9	2014 年	27	281	19	2	6	0
10	2014 年	27	690	14	3	10	0
11	2016 年	26	665	9	3	14	0
12	2015 年	26	534	14	2	8	2
13	2015 年	24	824	13	2	8	1
14	2014 年	23	1807	7	2	14	0
15	2015 年	22	376	11	3	7	1
16	2015 年	22	352	15	0	7	0
17	2014 年	22	314	10	3	9	0
18	2015 年	21	714	9	2	9	1
19	2014 年	21	213	12	3	6	0
20	2016 年	20	396	12	1	7	0
21	2016 年	20	679	12	1	7	0
22	2015 年	20	245	11	2	5	2
23	2014 年	20	290	13	2	5	0
24	2016 年	19	1131	13	0	6	0
25	2015 年	19	391	9	4	6	0

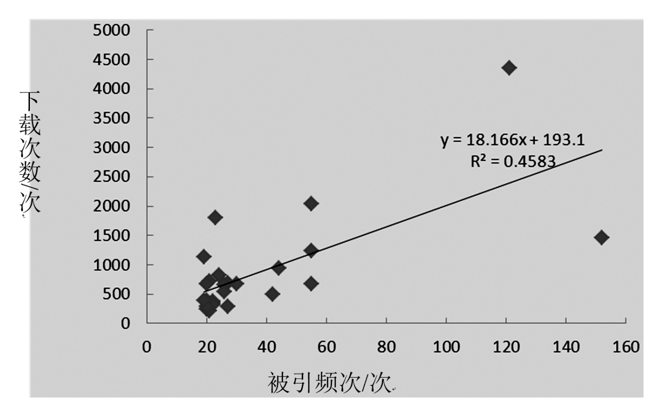


图 3　被引频次前 25 论文被引频次与下载频次相关性

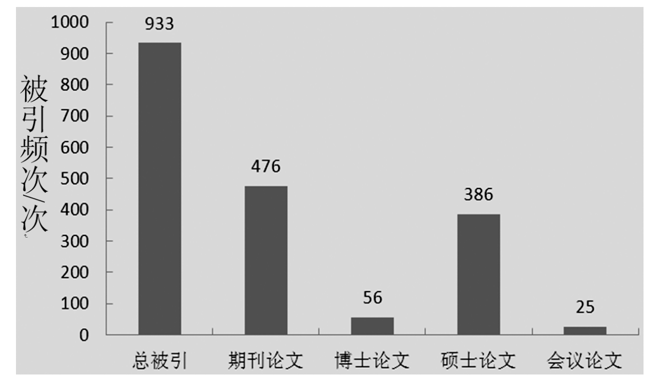


图 4　被引频次前 25 论文具体被引分散情况

具体被引分布情况见图 4。由图 2 可知：被引频次 TOP25 的论文总被引频次为 933，其中，期刊论文引用 476 次，占比 51.01%；博士论文引用 56 次，占比 6.00%；硕士论文引用 386 次，占比 41.10%；会议论文引用 25 次，占比 2.68%。

3. 分析与讨论

3.1 高下载频次和高被引频次论文发表年份不一致

被引频次是评价学术质量和影响力的重要指标， 从下载频次可以直接看出文献被读者使用的情况，反映了论文被关注的程度， 但是，不是所有的下载都会被引用。根据本文的研究结果， 文章出版后 3~5 年下载频次达到峰值，文章出版后 4~5 年被引频次达到峰值， 而高引用和高下载存在时间差解释了高下载和高被引论文年份不对应的现象。

3.2 下载频次和被引频次具有较好的相关性

论文下载频次越高，受读者关注的程度越高，被引用的几率就会增大，通过下载频次直接看出论文被读者使用的情况，视为对论文学术价值的先期反映。文献的被引频次与其质量高度相关，论文被引用意味着研究成果得到发展或评价，被引频次越高，效果越明显。

3.3 不能完全用下载频次对后续被引频次进行预测

人们通常认为论文的下载频次与被引频次正相关，即文章被下载得越多， 被引频次也就越高。下载一篇文章后是否引用它要看它的内在质量。当然， 下载频次与被引频次之间拥有某些相关性，但不是完全的线性相关，在一定的下载频次范围内（本文对应的下载频次为 500~1000 次）下载频次与被引频次是呈线性相关的，超过这个相关范围后被引频次和下载频次相关性不大,所以，不能完全用下载频次对后续的被引频次进行预测。媒

参考文献

[1] 温晓平，屈李纯，马秋明，等 . 西北农林科技大学学报（自然科学版）2001—2012 年高被引论文分析 [J]. 西北农林科技大学学报（自然科学版），2014，42（11）：225-234.

[2] 温晓平 . 入选综合性农业科学类中文核心期刊的 21 种农业类大学学报高被引论文学科分布的统计与分析 [J]. 农业图书情报学刊，2016，28（1）：51-56.

[3] 张晓丽，乐建鑫 . 高被引论文特征分析及启示 [J]. 东南大学学报（自然科学版），2012，23（6）：1008-1012.

[4] 杨弘 . 学术期刊被引频次与下载次数的关系 [J]. 安徽农业科学，2013，41（4）：1820-1821.

[5] 陆伟，钱坤，唐祥彬 . 文献下载频次与被引频次的相关性研究——以图书情报领域为例 [J]. 情报科学，2016，34（1）：1-6.

[6] 陈广仁，刘元珉 . 重视科技论文引用率，提高中国科技影响力 [J] . 科技导报（北京），2008， 26（5）：96-97.

[7] 丁佐奇，郑晓南 . 期刊影响因子， 论文被引证次数与学术质量评价的矛盾分析 [J]. 中国科技期刊研究，2009，20（2）：286-288.

[8] 李亚君 . 中文科技核心期刊高被引频次论文的统计分析 [J]. 河北理工大学学报（社会科学版），2010，10（4）：93-95.

[9] 郭强，赵瑾，刘思源，张芳，刘新新 . 科技论文下载次数的统计性质研究 [J]. 情报科学，2009，27（5）：690-694.

[10] Samad Jahandideh. Prediction of future citations of a research paper from number of its internet downloads[J].Medical Hypotheses，2007，69（2）：458-459.

[11] 郭强，赵瑾，刘新新 . 利用期刊下载次数估计后期被引频次的研究 [J]. 图书馆理论与实践，2010（11）：45-49.

[12] 张小强 . 期刊下载频次与被引频次及影响因子相关性——以中国知网 CSCD 与 CHSSCD 刊物为样本的计量分析 [J]. 情报理论与实践，2011，34（8）：36-40.

[13] Andrew J S.The top papers by download and citations from the International Journal of Cardiology in 2007[J].International Journal of Cardiology，2008（1）：1-3.

[14] 杜秀杰，赵大良，葛赵青，苗凌 . 学术论文的下载频率与被引频率的相关性分析 [J]. 编辑学报，2009，21（6）：551-553.

[15] 丁佐奇，郑晓南，吴晓明 . 科技论文被引频次与下载频次的相关性分析 [J]. 中国科技期刊研究，2010， 21（4）：467-470.

[16]Hauke Jan.Comparison of Values of Pearson' s And Spearman' s Correlation Coefficients on the Same Sets of Data [J].Quaestiones Geographicae，2011，30（2）：87-93.

（作者单位：《西南交通大学学报》编辑部）